DOCUMENT RESUME

ED 385 556 TM 023 978

AUTHOR Potenza, Maria T.; Stocking, Martha L.
TITLE Flawed Items in Computerized Adaptive Testing.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-6
PUB DATE Feb 94
NOTE 46p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; *Multiple Choice Tests; Scoring; Simulation; Test Bias; *Testing Problems; *Test Items; Thinking Skills; Verbal Tests
IDENTIFIERS *Flawed Items; Test Repeaters; Test Rescoring

ABSTRACT

A multiple choice test item is identified as flawed if it has no single best answer. In spite of extensive quality control procedures, the administration of flawed items to test-takers is inevitable. Common strategies for dealing with flawed items in conventional testing, grounded in the principle of fairness to test-takers, are reexamined in the context of adaptive testing. These are usually removing the flawed item or rescoring it in a reasonable fashion. An additional strategy, available for adaptive testing, of retesting from a pool cleansed of flawed items, was compared to the existing strategies. A simulation was performed for 1,300 simulees from a uniform distribution of proficiency on a test of verbal reasoning. Results were weighted to reflect the results of a typical distribution of proficiency. Retesting was found to be no practical improvement over current strategies. Six tables present analysis details, and an appendix explains the weighted derivations. (Contains 14 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-94-6

ED 385 556

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

FLAWED ITEMS IN COMPUTERIZED ADAPTIVE TESTING

Maria T. Potenza
Martha L. Stocking

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
February 1994

TM023978

FLAWED ITEMS IN COMPUTERIZED ADAPTIVE TESTING[1,2]

Maria T. Potenza
University of Nebraska
Lincoln, Nebraska

Martha L. Stocking[3]
Educational Testing Service
Princeton, New Jersey

---

[1] This research was supported in part by the Program Research Planning Council of Educational Testing Service.

[2] The authors gratefully acknowledge the advice and help of Dr. Charles Lewis for some of the analyses, and particularly for the Appendix.

[3] The authors' names are listed alphabetically.

Copyright © 1994. Educational Testing Service. All rights reserved

Abstract

A (multiple-choice) test item is identified as flawed if it has no single best answer. In spite of extensive quality control procedures, the administration of flawed items to test-takers is inevitable. Common strategies for dealing with flawed items in conventional testing, grounded in the principle of fairness to test-takers, are reexamined in the context of adaptive testing. An additional strategy, available for adaptive testing, of retesting from a pool cleansed of flawed items, is compared to the existing strategies. Retesting was found to be no practical improvement over current strategies.

Key Words: computerized adaptive testing, flawed items, monte carlo simulations.

Introduction

Large testing organizations produce thousands of new items every year. These items typically are reviewed and revised many times, by content experts, test specialists, and sensitivity reviewers, before being presented to test-takers. An item that has survived this extensive review process is then usually 'pretested', that is, included with other such items and administered to test-takers but not included in test scores. The purpose of this final step is to identify items with appropriate statistical properties. At every stage in this extensive development process, items may be discarded as deficient in one or more features that are associated with good test items.

Occasionally, in spite of the care taken in the development of items that count towards test-takers' scores, a (multiple-choice) item will be identified as 'flawed' when it appears in a test, that is, the item has no single best answer. Testing organizations have developed various strategies in the context of conventional (linear) paper-and-pencil testing for dealing with the discovery of flawed items that were originally intended to count towards test-takers' scores. The professional principle underlying such strategies is fairness to test-takers, in conformance with the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1985).

Recent advances in psychometrics and computing technology have led to the development of testing paradigms that are very different from linear paper-and-pencil testing, such as computerized adaptive testing (CAT) or computerized mastery testing (CMT), for example, Eignor, Way, Stocking, & Steffen (1993), Lord (1977), Schaeffer, Steffen, & Golub-Smith (1993), Sheehan

and Lewis (1992), Stocking and Swanson (1993), and Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg & Thissen (1990). Existing strategies for handling flawed items must be reexamined for appropriateness in these new testing paradigms, and new strategies may be required. In this paper, we will discuss current strategies, their applicability in the context of computerized adaptive testing (CAT), and a monte carlo experiment to evaluate various potential strategies in adaptive testing.

Typical Flaws and Current Strategies

Conventional Testing

Conventional linear paper-and-pencil tests are typically administered relatively few times a year to large numbers of test-takers simultaneously. There is a lapse in time between testing and the reporting of test scores to individuals and institutions while answer sheets are collected in a central location and translated into computer-readable records, tests are scored and equated, and score reports are produced and mailed. These characteristics of conventional testing have facilitated the development of certain strategies for dealing with flawed items. First, because of the time lapse for score reporting, some actions can be taken before score reporting ever occurs. Second, because it is easy to identify all test-takers who were administered the flawed item, it is also easy to rereport scores within the time the scores are still considered meaningful.

Strategies for dealing with flawed items fall into two major categories: either remove the flawed item from the test or rescore the flawed item in a reasonable fashion. In either case test scores are reported or possibly

rereported for affected test-takers and a particular strategy is chosen in light of what is fairest for all test-takers.

Typical flaws include the following:

1) No correct answer

An item can become obsolete or incorrect because of societal changes, scientific discoveries, and so forth. Or an error that changes the meaning of the item sufficiently so that no answer is correct can be introduced inadvertently into item text.

2) Multiple correct answers

It is possible that a test-taker with a novel point of view may discover that from a particular perspective an item has a different right answer or multiple right answers, as in, for example, Wainer (1983).

3) An incorrect scoring guide or key was used

For this type of flaw, the item itself is valid, but incorrect information about correct answers was used in the process of machine scoring answer sheets.

Table 1 shows which strategies are most frequently appropriate for the various types of flaws before scores are rereported to test-takers and institutions.

Insert Table 1 about here

Adaptive Testing

In the administration of a conventional test, every test-taker responds to the same set of items. In adaptive testing, where an item is selected based on responses to previous items, it is theoretically possible for every

test-taker to receive a different test. In addition, since adaptive tests are administered on a computer, testing is nearly continuous and score reporting can be immediate. These two characteristics make the strategies outlined above more difficult to implement since identification of test-takers who received a flawed item is more complex and scores may have already been reported. While this serves to make the book-keeping task more difficult, and may always require rereporting of scores, it does not obviate any of the strategies outlined above that are used to implement the principle of fairness to test-takers.

Adaptive testing presents an additional problem that is not present in conventional testing. Responses to each item contribute to a test-taker's score, as in conventional testing. However, responses to each item also determine which items are selected subsequent to a given item. Thus there is the potential that a flawed item might lead to the routing of a test-taker through the pool of items in such a way as to unfairly influence his or her final score. Simply rescoring or removing a flawed item, as is done in conventional testing, may not be sufficient to compensate for the full effects of a flawed item in adaptive testing.

An effective strategy for dealing with flawed items in adaptive testing, then, might be to remove a flawed item from the item pool, and offer test-takers the opportunity to repeat the test without cost to them. This is a costly alternative to rescoring and rereporting, both in terms of the actual costs of adaptive test administration and in terms of inconvenience to test-takers. In circumstances where the difference in scores between rescoring and a second adaptive test from a reduced pool are comparable to the difference expected from two adaptive tests from the same reduced item pool, retesting

might prove to be unnecessary. The monte carlo experiment is designed to investigate these issues.

The Adaptive Test

The particular adaptive test chosen for this experiment is a test designed to measure verbal reasoning in a high-stakes admissions testing context. The verbal measure was chosen over the other two measures available because this measure represents a balance of discrete items as well as items associated with reading passages, whereas one of the other two measures available consisted of predominantly discrete items (for which one would expect flawed items to have a smaller impact on routing) and the other consisted of predominantly set-based items (for which one would expect flawed items to have a larger impact on routing). For the companion linear paper-and-pencil testing program in the last calendar year, approximately .01% of the items across all tests were identified as flawed.

The psychometrics underlying the adaptive test are based on the three parameter logistic Item Response Theory (IRT) model (Lord, 1980). The item pool consists of 381 items and passages that are identified along 38 different (usually nonmutually exclusive) features associated with subject matter, item type, and so forth. The items were calibrated and placed on the same metric using the computer program LOGIST (Wingersky, 1983). The item selection in the adaptive test employs the methodology of the weighted deviations model of Stocking and Swanson (1993) with the extended Sympson and Hetter (1985) exposure control methodology (Stocking, 1992) to increase item security. (For details of the test design process, see Eignor, et al., 1993).

In the weighted deviations approach to adaptive testing, item properties or features are taken into account along with statistical properties in the selection of items. This is to insure that each adaptive test produced from the pool matches a set of test content and item type specifications and is therefore as parallel as possible to any other test in terms of content and type of items, while being tailored to an individual examinee in terms of difficulty. The weighted deviations approach also allows specification of overlapping items that cannot be administered in the same adaptive test. In addition, it is possible to restrict item selection to blocks of items, either because they are associated with a common stimulus or common directions or any other feature that test specialists deem important.

In summary, in the weighted deviations model, the next item selected for administration is the item that

1) is the most informative item possible at a test-taker's estimated ability level, while
2) simultaneously contributing the most to the satisfaction of all other constraints in addition to the constraints on item information.

At the same time, it is required that the item

3) does not appear in an overlap group containing an item already administered, and
4) is in the current block (if the last item was in a block), starts a new block, or is in no block.

The Sympson and Hetter exposure control methodology further restricts item selection by determining if the selected item is likely to be overexposed if administered, based on exposure control parameters developed over a series of simulations with a (simulated) typical group of test-takers. If so, this

methodology forces the administration of an item that has been administered less frequently. For this adaptive test, the maximum observed exposure of an item is about .24, meaning that no more than 24% of a typical group of test-takers will receive the most popular item in the pool. The estimated reliability, computed using Green, Bock, Humphreys, Linn, & Reckase (1984, equation 6) of the adaptive verbal measure at the end of the test design simulations was .902.

The adaptive test is scored by converting the final (maximum likelihood) estimate of examinee proficiency to an estimated number right true score on a (linear) 76-item reference test that was previously scaled to the score reporting metric. For this test, the raw (estimated number right) scores range from a chance level of 13 to a high of 76.

The Monte Carlo Experiment

The Number of Flawed Items

The starting point for this experiment is the final simulation to establish the test design for the adaptive verbal measure. This baseline simulation was performed for 1300 simulees from a uniform distribution of proficiency and the results were weighted to reflect the results for a typical distribution of proficiency. The typical distribution of proficiency was obtained using the methods of Mislevy (1984). All subsequent simulations required by the current experiment were performed in a similar fashion.

In all, five experimental conditions were considered:

1) Twenty-five most popular

To simulate a worst case, the 25 most popular (that is, most frequently administered) items in the baseline simulation were

considered to be flawed. Thus nearly every simulee should receive at least one flawed item.

More realistic conditions would dictate that there might be approximately .04 flawed items in a 381-item pool (.01% of 381). We chose to model the substantially larger number of two flawed items for the remaining conditions as a conservative approach that would facilitate the comparison of the various conditions. The remaining four conditions are as follows:

2) Two most popular

The second condition considered the two most popular items in the baseline simulation to be flawed. A substantial number of simulees can be expected to receive at least one.

3) Two typical items

The third condition considered two items with average exposure rates from the baseline simulation to be flawed. This is probably the most realistic condition in terms of the exposure rate of items.

4) Two most popular as first items

The fourth condition considered as flawed those two items that appeared most frequently as the first item in the baseline simulation. In this condition, one would expect to find the biggest impact on the routing of the simulee through the remainder of the pool.

5) Two most popular as last items

The fifth condition considers as flawed those two items that appeared most frequently as last items in the baseline simulation.

BEST COPY AVAILABLE

These items, of course, can have no effect on the adaptive test routing, but may have an effect on final test scores.

Modeling Flawed Items

Because adaptive testing is based in Item Response Theory, monte carlo simulations are possible. In this context, right and wrong responses can be generated for simulated examinees (simulees) in conformance with the item response model chosen and the estimated item parameters for each item (Lord, 1980, Hambleton and Rovinelli, 1973). In typical simulations of adaptive testing, estimates of item parameters are obtained from pretesting the items, and the estimates are treated as if they were true values in the generation of simulee responses and in the routing of simulees through the adaptive test.

The simulation of flawed items requires a slightly different philosophical approach. We use the estimated item parameters as true values for the selection of items in the adaptive test. However, right and wrong responses for simulees are generated using a different set of item parameters that reflect the fact that the item is flawed when it is administered in the context of counting towards a test score, but not flawed when it was pretested. (If it were identified as flawed on the basis of pretest data, it would not have been included in the item pool).

For example, suppose that a particular item was pretested and determined at that time to be an appropriate item. Suppose that somehow the text of the item became corrupted over time so that when the item is used in an adaptive test, it has no correct answer. The item parameters estimated from pretesting are used by the item selection algorithm in the selection of items. However, a test-taker sees the item and realizes that there is no correct answer. Thus

the item appears impossibly difficult to the test-taker and the test-taker's response is modeled by a second set of item parameters.

For purposes of this experiment, two different kinds of flawed items were simulated in the following conservative approach, in order to assess the effects of circumstances more extreme than are likely to be found in actual practice. For flawed items for which there is no correct answer when presented to test-takers, we assumed that the item would become very difficult and all simulees would respond incorrectly. For this situation, simulee responses were generated to a very highly discriminating item that was very difficult, and impossible to answer correctly by guessing (a = 3, b = 10, and c = 0). For flawed items with more than one correct answer or for which the incorrect scoring guide was used, we assumed that the item would also become very difficult but that some simulees would respond correctly by chance alone. For this situation, simulee responses were generated for item parameters of a = 3, b = 10, and c = .25.

In assigning flaws to items, a simple pattern of alternation was followed. In the first condition with 25 flawed items, items 1, 3, 5, ...25, were assigned a=3, b=10, and c=0 as parameters for generating simulee responses, for a total of 13 such items. Items 2, 4, 6, ... 24, were assigned a=3, b=10, and c=.25 as parameters for generating simulee responses for a total of 12 such items. In the other four conditions with two flawed items, the first item was always identified with the first type of flaw; the second item was always identified with the second type of flaw.

Methods of Rescoring

As seen in Table 1, there are a number of possible strategies that may be used for different kinds of detected flaws in items. We chose to compare

four different strategies. The first three strategies ignore the type of flaw simulated for the item. In the first strategy we remove flawed items from simulees' response strings and rescore the adaptive test based on the reduced set of responses. A second strategy is to score any answer correct. In the context of a simulation, this results in changing all incorrect answers to correct answers and rescoring the adaptive test. A third strategy is to rescore flawed items with the correct key. This is accomplished by generating a new response for a flawed item based on the estimated item parameters obtained from pretesting and rescoring the adaptive test.

In the fourth and final strategy, we take into account the type of flaw being simulated in an item. Items simulated as flawed because they have no correct answer are removed from the response string before rescoring. For items simulated as flawed because of more than one correct answer or because of an incorrect scoring guide, new responses are generated using the pretest item parameters and the response string is rescored.

The Reduced Pools

The final alternative for dealing with flawed items in the context of adaptive testing is to remove flawed items from the pool, and to offer retesting from the reduced pool for those test-takers who received flawed items from the original pool. In order to simulate the results of this alternative, the original pool was reduced five separate times in parallel with the five conditions studied and the simulations were repeated.

The Results

The Reduced Pools

Before any useful comparisons can be made, it is necessary to examine the consequences of reducing the original baseline pool five separate times in order to insure that the results are in conformance with what would actually happen in practice. Table 2 displays the maximum observed exposure rates for an initial adaptive test simulation on each of the five reduced pools. As can be seen from this Table, three of the pools produced maximum exposure rates that were in excess of what was considered desirable for the baseline pool. This was anticipated for the first condition in which the 25 most popular items were removed from the pool. It was not anticipated for the remaining conditions in which just two items were removed from the pool. However, it is clear from the table that which two items are removed can have differential effects on observed maximum exposure rates, with the removal of the two most popular items having an effect similar to removing the 25 most popular items.

Therefore, additional extended Sympson and Hetter iterations were performed for the three pools requiring such iterations in order to adjust the exposure control parameters to take into account the new pool sizes. The maximum exposure rates for the adjusted exposure control parameters are given in parentheses in Table 2, and seems satisfactory.

Insert Table 2 about here

Rescoring Methods

Tables 3a through 3e display aspects of the simulations and the rescoring. The first column in each table gives the reliability and test length of the baseline unflawed simulation and is the same for all five tables. The second column gives the same information for the simulation in which the items were considered flawed and the simulees' responses to flawed items were generated using the alternate set of item parameters. The next four columns give the results for each rescoring of the flawed simulation, and the final column gives the reliability and test length for the simulation on the reduced pool after the new extended Sympson and Hetter iterations were performed (if required).

Insert Tables 3a-3e about here

For each rescoring, the Tables display the mean score difference (rescored simulation minus flawed simulation results) for simulees receiving flawed items in a typical population of test-takers. Thus if the 25 most popular items are simulated as flawed and then removed from scoring, the average test-taker score increases 5.06 raw score points. If the flawed items are rescored as all correct, the increase is 7.27. If the correct key is used to rescore items, the increase is 5.06, and if the items are rescored taking into account the type of flaw in the item, the average increase is 5.03.

Tables 3a through 3e also display the proportion of a typical population that could be expected to have at least one flawed item. If the 25 most popular items are simulated as flawed, 100% of this typical distribution can be expected to have at least one flawed item, while if only the two most

popular items are simulated as flawed, 40% of a typical population have at least one flawed item. If two typical items are simulated as flawed, only about 21% of a typical group of test-takers receives one or more flawed items.

As expected, an adaptive test with flawed items has lower reliability than an adaptive test without flawed items from an item pool of the same size. This reduction in reliability is largest when the number of flawed items is greatest (from .902 to .757 for Table 3a). However, it is also fairly large if the two items that are simulated as flawed are the two that have the most impact in the routing through the pool because they are the two items appearing most frequently in the first position of the adaptive test simulation on the baseline pool (Table 3d). The reliability is substantially improved by any method of rescoring, and also by removing the items from the pool and repeating the testing. However, the reliability of the adaptive test from the reduced pool is usually slightly lower than that for the baseline pool, as expected since the reduced pool contains fewer items.

Rescoring flawed items by accepting any response as a correct answer results in the largest score increase -- sometimes double that of the other rescoring methods. The other three methods of rescoring result in mean score increases that are very similar to each other. For the conditions in which only two items are simulated as flawed, this is not surprising. The effect of removing two items from a 30-item test, or generating two new responses with the right item parameters, or removing one item and generating a new response for the other should raise test scores slightly because each method is equivalent to discarding two very hard items and substituting either no items, or items that are easier. The effects of rescoring when two items are changed

are very similar because two items is a small percentage of the 30 items on which the maximum likelihood estimate of proficiency is based.

For the condition in which 25 items are simulated as flawed, one might expect a greater difference among the scoring methods. However, a detailed examination of the conditional distribution of the number of flawed items in an adaptive test (conditional on true ability) reveals that the number of flawed items per simulee is usually quite small. Only two simulees out of 1300 received 10 or more flawed items. Thus the same argument -- that only a small percentage of items for any individual simulee is flawed -- holds.

For the four conditions involving only two flawed items, if the two items appear most frequently in the first position of an adaptive test, the mean score differences of all rescoring methods are greater by roughly a factor of five than those for two flawed items that appear most frequently in the last position of an adaptive test. The most popular first items in an adaptive test are likely to be informative (psychometrically) and most appropriate for test-takers of typical proficiency. The most popular last items in an adaptive test are likely to be less informative but still appropriate for test-takers of typical proficiency. Rescoring of first items has a larger effect than rescoring of last items due to the differential impact on (maximum likelihood) scoring of the more informative items and the less informative items.

The effects for two typical items and for the two most popular items are between these two extremes. The most realistic situation is likely to be that in which the two flawed items are items with typical exposure rates. In this situation, rescoring methods can be expected to result in about a half-point to one point increase in average test scores. The most appropriate rescoring

method, which takes into account the type of flawed item, results in a mean score increase for a typical population of .6 of a raw score point.

Rescoring vs. Retesting from a Reduced Pool

The above results illuminate the differences among four different methods of rescoring an adaptive test. A key question is how different any method is from offering test-takers who received flawed items a second adaptive test from a pool from which the flawed items have been removed. For this analysis, we considered only the appropriate method of rescoring that takes into account the nature of the flawed items.

Table 4 displays the root mean squared score differences (RMSDs) between various simulations of interest for each of the five conditions **for those simulees who received flawed items**. Formulae for the computation of these are given in the Appendix. Column 1 contains the RMSDs that can be expected from two administrations of CAT from the baseline pool for simulees who received flawed items in each of the five conditions. These numbers differ from each other only because the conditional (on true ability) distributions of simulees receiving flawed items vary from condition to condition, with all simulees in the first condition receiving at least one flawed item.

Insert Table 4 about here

The second column contains the expected RMSDs (for simulees receiving flawed items) between a rescored CAT and a CAT from the reduced pool. The third column contains the expected RMSDs (for simulees receiving flawed items) between two CATs administered from the reduced item pool.

The RMSDs between two CATs from the baseline pool (column 1) are typically smaller than the corresponding from the RMSDs from the reduced pool (column 3) for most conditions. This is to be expected since the reduced pool is smaller than the baseline pool and therefore item selection is less optimum, causing greater variability upon retesting from the same pool.

The RMSDs between a rescored CAT and a CAT from the reduced pool are larger than the RMSDs between two CATs from the reduced pool for the first, second, and fourth conditions. These three conditions can be expected to have the most effect on routing in the simulation of flawed items. For the other two conditions with no effect (the fifth condition) or random effects (the third condition) on routing the direction of the differences between RMSDs is reversed. Although all differences between RMSDs are small (the maximum is on the order of .5 a raw score point for the 25 flawed item simulation) and these results could be due to sampling error, they are somewhat disquieting. The RMSD between two replications from the same pool is related to test-retest reliability, and this comparison is akin to finding that the correlation between scores on two different measures is higher than the correlation between (repeated) scores on the same measure. This is further illustrated in Table 5, which gives the correlation between pairs of scores (for simulees with flawed items) for all the conditions.

Insert Table 5 about here

To further investigate this result, we chose to analyze in more detail one of the anomalous conditions -- two typical items simulated as flawed. Table 6 shows the results of seven additional replications for this condition

in addition to the original shown in Table 4, along with the means and standard deviations of the RMSDs for all replications.

Insert Table 6 about here

The difference of the mean RMSD from two administrations from the baseline pool (column 1) minus the mean RMSD from two administrations from the reduced pool (column 3) is -.117, with a standard error of .004. Thus the 95% confidence interval for the difference is [-.128, -.106], which does not include zero. This in comforting since we expect that the RMSD for a smaller pool should be larger than for larger pool. The difference of the mean RMSD from a rescored test and a retesting with a reduced pool (column 2) minus the mean RMSD from two administrations from the reduced pool (column 3) is -.139 with a standard error of .079, giving a 95% confidence interval for the difference of [-.326, .049]. Since this confidence interval includes zero, we can view with more certainty the apparently anomalous results in Tables 4 and 5 as consequences of sampling error.

Discussion

In spite of quality control procedures followed by testing organizations, the presentation to test-takers of flawed items that are originally intended to count toward test scores is inevitable. Testing organizations have already established various strategies for dealing with flawed items in conventional (linear) tests, once they are discovered. The principle underlying all such strategies is fairness to test-takers.

The purpose of the current effort was twofold: to investigate the applicability of currently known strategies in the context of adaptive testing, and to compare the current strategies with an additional counterpart in adaptive testing of offering retesting from a reduced adaptive testing pool from which flawed items have been removed.

Conventional strategies work well with the adaptive test chosen for this study. Accepting any response as correct increases the average test score more than simply removing flawed items, rescoring flawed items with correct answer keys, or tailoring the strategy to the nature of the flawed items, as would be expected. The magnitude of the increases, based on the simulated results presented here, depends upon a number of factors. If a large number of frequently administered items are simulated as flawed, the mean score increase for any method of rescoring is larger than if the number of flawed items is small. If the number of items is small (but still an order of magnitude larger than would be found in actual practice) the magnitude of the mean score increase for any rescoring method depends upon the location of the flawed items in the adaptive test. The largest impact is found for flawed items that are frequently the first item in adaptive tests from the pool, while the smallest impact is found for such items when they are most frequently administered last. The mean score differences for typical items, as well as frequently administered items are between these two extremes.

In terms of fairness to test-takers, does it make any difference whether an adaptive test is simply rescored or if they take another test from a reduced pool? The practical answer is "no". In the worst case, in which a large number of frequently administered items are flawed, or if the number of flawed items is small but can be expected to have an impact on routing, the

root mean squared difference between a rescored CAT and a CAT from a reduced pool is larger than that for two testings from the same reduced pool. However, the largest observed difference between RMSDs, for the worst case, was only about one-half a raw score point; for the other two cases in which routing can be expected to be influenced, the differences between RMSDs was on the order of one-tenth a raw score point. For the two cases in which the number of items is small and can be expected to have little or no influence on routing through an item pool, the differences between RMSDs were negligible.

Whether the results of this simulation study will generalize to other adaptive tests is, of course, not known. However, the adaptive test chosen for this study is fairly typical of adaptive tests being prepared for large scale implementation in the near future, and the various rescoring strategies studied are typical of those most frequently used with conventional linear testing. Thus the prospects for generalization appear to be good, although this should be confirmed with additional studies using different adaptive tests.

Appendix[1]

The Weights

We have $g(\theta) = g(\xi)$, the distribution of proficiency in a typical group of examinees, from the method of Mislevy (1984) at K discrete values of (nearly) equally spaced ξ_k, where ξ_k is the number correct true score on the reference set of items used for scoring purposes. We wish to make comparisons of various root mean squared differences only for those simulees who received flawed items in a particular simulation. We need $\hat{f}(\xi|\text{flawed item})$, which we can obtain by Bayes theorem as follows.

Let $g(\xi_k)$ be the original weights, or the prior, $k = 1, \ldots, K$. Let $\hat{P}(\text{flawed item}|\xi_k)$ $k = 1, \ldots, K$, be the sample proportion of simulees receiving at least one flawed item, given true score. This information is available from a simulation of flawed items. Then the estimated posterior probability

$$\hat{f}(\hat{\xi}_k) = \hat{f}(\hat{\xi}_k|\text{flawed item}) = \frac{\hat{P}(\text{flawed item}|\xi)\, g(\xi_k)}{\sum_{k'} \hat{P}(\text{flawed item}|\xi_{k'})\, g(\xi_{k'})} \quad . \tag{A1}$$

These new weights will be used in the computation of various root mean squared differences.

The Expected Squared Score Difference of Two CATs from the same Pool

Suppose a person with true score ξ takes an adaptive test from an item pool and receives $\hat{\xi}_1$, as a test score. Suppose the same person takes a second adaptive test from the same item pool and receives $\hat{\xi}_2$ as the second

---

[1] The derivations in this Appendix and the computer programs to obtain the actual quantities are due to Dr. Charles Lewis. The authors are extremely grateful for his help, interest, and advice.

test score. Neither true ability nor the item pool changes between the two testings. We want

$$E\left\{E\left[\hat{\xi}_1 - \hat{\xi}_2\right]^2 | \xi\right\}, \tag{A2}$$

the expected squared score difference of two CATs from the same pool. The interior expectation is the within group expectation, and the exterior expectation is the expectation over the population of simulees who received flawed items. For convenience, drop the conditional notation for a moment, and also use the notation $E(\hat{\xi}_1|\xi) = \mu_1$, $\text{var}(\hat{\xi}_1|\xi)$, $E(\hat{\xi}_2|\xi) = \mu_2$, and $\text{var}(\hat{\xi}_2|\xi)$. Now

$$\begin{aligned} E\left[\hat{\xi}_1 - \hat{\xi}_2\right]^2 &= E\left[\left(\hat{\xi}_1 - \mu_1\right) - \left(\hat{\xi}_2 - \mu_2\right) + \left(\mu_1 - \mu_2\right)\right]^2 \\ &= E\left[\left(\hat{\xi}_1 - \mu_1\right)^2\right] + E\left[\left(\hat{\xi}_2 - \mu_2\right)^2\right] + \left(\mu_1 - \mu_2\right)^2 , \end{aligned}$$

since test scores are uncorrelated when ξ is fixed. Using the notation above, and re-introducing the conditional notation,

$$E\left[\left(\hat{\xi}_1 - \hat{\xi}_2\right)^2 | \xi\right] = \text{var}(\hat{\xi}_1|\xi) + \text{var}(\hat{\xi}_2|\xi) + \left[E(\hat{\xi}_1|\xi) - E(\hat{\xi}_2|\xi)\right]^2 . \tag{A3}$$

However, except for sampling error

$$\text{var}(\hat{\xi}_1|\xi) = \text{var}(\hat{\xi}_2|\xi)$$

and

$$E(\hat{\xi}_1|\xi) = E(\hat{\xi}_2|\xi),$$

then

$$E\left[\left(\hat{\xi}_1 - \hat{\xi}_2\right)^2 | \xi\right] = 2\ \text{var}(\hat{\xi}_1|\xi).$$

To get the desired population expectation,

$$E\left[E\left(\hat{\xi}_1 - \hat{\xi}_2\right)^2 | \xi\right] = E\left[2\ \mathrm{var}\left(\hat{\xi}_1 | \xi\right)\right] = 2\ E\left[\mathrm{var}\left(\hat{\xi}_1 | \xi\right)\right] . \tag{A4}$$

The square root of this quantity is the standard deviation of the distribution of difference scores between the two administrations.

The Expected Squared Score Difference of Two CATs from Different Pools

Suppose a person with true score ξ takes an adaptive test from an item pool and receives $\hat{\xi}_1$ as a test score. This can be a rescored adaptive test from a pool with flawed items. Suppose the same person takes a second adaptive test from a reduced pool and receives $\hat{\xi}_2$ as a test score. True ability has not changed, but the item pool is now different. As before, we want the expected squared score difference given in (A2). Since these scores are also uncorrelated when ξ is fixed, we can begin with equation (A3) (the interior expectation in (A2)). However, in contrast to the previous situation, the conditional means and variances will not be equal since the item pools are now different so the equation is more complex. Thus we want

$$E\left\{E\left[\left(\hat{\xi}_1 - \hat{\xi}_2\right)^2 | \xi\right]\right\} = E\left[\mathrm{var}\left(\hat{\xi}_1 | \xi\right)\right] + E\left[\mathrm{var}\left(\hat{\xi}_2 | \xi\right)\right] = E\left\{\left[E\left(\hat{\xi}_1 | \xi\right) - E\left(\hat{\xi}_2 | \xi\right)\right]^2\right\} \tag{A5}$$

Sample Estimates

Thus far we have derived population expressions of various means, mean squares and so forth. The obvious corresponding sample expressions are not

always unbiased. Using a slightly different but more general notation in this section, we derive unbiased estimates of the necessary population quantitites.

Suppose we have x_i, $i = 1, \ldots, K$, with the x_i identically distributed with mean μ_i and covariance $\frac{\sigma_i^2}{n_i}$. Also we have weights w_i, $i = 1, \ldots, K$ with all $w_i \geq 0$ and $\Sigma w_i = 1$. Define $\bar{x} = \Sigma w_i x_i$ and $\bar{\mu} = \Sigma w_i \mu_i$. We want to use $\Sigma w_i (x_i - \bar{x})^2$ to estimate $\Sigma w_i (\mu_i - \bar{\mu})^2$.
We have

$$E\left\{\Sigma w_i (x_i - \bar{x})^2\right\} = E\left\{\Sigma w_i \left[(x_i - \mu_i) + (\mu_i - \bar{\mu}) + (\bar{\mu} - \bar{x})\right]^2\right\},$$

which, after some algebra, is equivalent to

$$= \Sigma w_i \frac{\sigma_i^2}{n_i} + \Sigma w_i (\mu_i - \bar{\mu})^2 - \mathrm{var}(\bar{x})$$

Now $\mathrm{var}(\bar{x}) = \mathrm{var}(\Sigma w_i x_i) = \Sigma w_i^2 \frac{\sigma_i^2}{n_i}$.

So

$$E\left\{\Sigma w_i (x_i - \bar{x})^2\right\} = \Sigma w_i (\mu_i - \bar{\mu})^2 + \Sigma w_i \frac{\sigma_i^2}{n_i} - \Sigma w_i^2 \frac{\sigma_i^2}{n_i}.$$

Thus an unbiased estimate of $\Sigma w_i (\mu_i - \bar{\mu})^2$ would be

$$\Sigma w_i (x_i - \bar{x})^2 - \Sigma w_i (1 - w_i) \frac{\hat{\sigma}_i^2}{n_i}.$$

Now consider two sets of variables, x_i and y_i, all mutually independent,

with $E(x_i) = \mu_{x_i}$, $E(y_i) = \mu_{y_i}$, $\text{var}(x_i) = \frac{\sigma^2_{x_i}}{n_{x_i}}$, $\text{var}(y_i) = \frac{\sigma^2_{y_i}}{n_{y_i}}$. Again we have nonnegative weights w_i with $\Sigma w_i = 1$. Define $\bar{x} = \Sigma w_i x_i$ and $\bar{y} = \Sigma w_i y_i$, $\bar{\mu}_x = \Sigma w_i \mu_{x_i}$, $\bar{\mu}_y = \Sigma w_i \mu_{y_i}$. Consider $\Sigma w_i (x_i - \bar{x})(y_i - \bar{y})$ as an estimate of $\Sigma w_i (\mu_{x_i} - \bar{\mu}_x)(\mu_{y_i} - \bar{\mu}_y)$.

Now

$$E\left\{\Sigma w_i (x_i - \bar{x})(y_i - \bar{y})\right\} = \Sigma\left[w_i\, E(x_i - \bar{x})\, E(y_i - \bar{y})\right]$$
$$= \Sigma w_i (\mu_{x_i} - \bar{\mu}_x)(\mu_{y_i} - \bar{\mu}_y) ,$$

because of the mutual independence of x_i and y_i. Therefore, $\Sigma w_i (x_i - \bar{x})(y_i - \bar{y})$ is an unbiased estimator of $\Sigma w_i (\mu_{x_i} - \bar{\mu}_x)(\mu_{y_i} - \bar{\mu}_y)$.

Next, consider $(\bar{x} - \bar{y})^2$ as an estimate of $(\bar{\mu}_x - \bar{\mu}_y)^2$. Taking expectations again, we have

$$E\left\{(\bar{x} - \bar{y})^2\right\} = E\left\{\left[(\bar{x} - \bar{\mu}_x) + (\bar{\mu}_x - \bar{\mu}_y) + (\bar{\mu}_y - \bar{y})\right]^2\right\} ,$$

which can be shown to be equal to

$$= \Sigma w_i^2 \frac{\sigma^2_{x_i}}{n_{x_i}} + (\bar{\mu}_x - \bar{\mu}_y)^2 + \Sigma w_i^2 \frac{\sigma^2_{y_i}}{n_{y_i}} .$$

So an unbiased estimator of $(\bar{\mu}_x - \bar{\mu}_y)^2$ would be

$$(\bar{x} - \bar{y})^2 - \Sigma w_i^2 \frac{\hat{\sigma}^2_{x_i}}{n_{x_i}} - \Sigma w_i^2 \frac{\hat{\sigma}^2_{y_i}}{n_{y_i}} .$$

Finally, consider $\Sigma w_i (\bar{x}_i - \bar{y}_i)^2$ as an estimator for $\Sigma w_i (\mu_{x_i} - \mu_{y_i})^2$. Taking

expectations we have

$$E\left\{\Sigma w_i(\bar{x}_i - \bar{y}_i)^2\right\} = \Sigma w_i\, E\left\{(\bar{x}_i - \bar{y}_i)^2\right\} .$$

Now

$$\begin{aligned} E\left\{(\bar{x}_i - \bar{y}_i)^2\right\} &= E\left\{\left[\left[(\bar{x}_i - \bar{y}_i) - (\mu_{x_i} - \mu_{y_i})\right] + (\mu_{x_i} - \mu_{y_i})\right]^2\right\} \\ &= \operatorname{var}(\bar{x}_i - \bar{y}_i) + (\mu_{x_i} - \mu_{y_i})^2 \\ &= \frac{\sigma^2_{x_i}}{n_{x_i}} + \frac{\sigma^2_{y_i}}{n_{y_i}} + (\mu_{x_i} - \mu_{y_i})^2 . \end{aligned}$$

So

$$E\left\{\Sigma w_i(\bar{x}_i - \bar{y}_i)^2\right\} = \Sigma w_i \frac{\sigma^2_{x_i}}{n_{x_i}} + \Sigma w_i \frac{\sigma^2_{y_i}}{n_{y_i}} + \Sigma w_i(\mu_{x_i} - \mu_{y_i})^2 .$$

Therefore $\Sigma w_i(\bar{x}_i - \bar{y}_i)^2 - \left(\Sigma w_i \frac{\hat{\sigma}^2_{x_i}}{n_{x_i}} + \Sigma w_i \frac{\hat{\sigma}^2_{y_i}}{n_{y_i}}\right)$ is an unbiased estimator of

$\Sigma w_i(\mu_{x_i} - \mu_{y_i})^2$.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Eignor, D. R., Way, W. D., Stocking, M. L., and Steffen, M. (1993). Case studies in computer adaptive test design through simulation (Research Report 93-56). Princeton, NJ: Educational Testing Service

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Hambleton, R. K., and Rovinelli, R. (1973). A Fortran IV program for generating examinee response data from logistic test models. Behavioral Science, 17, 73-74.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49, 359-381.

Schaeffer, G., Steffen, M., Golub-Smith, M. (1993). Introduction of a Computer Adaptive GRE General Test. (Research Report XX-XX). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1992). Controlling Item Exposure Rates in a Realistic Adaptive Testing Paradigm. (Research Report 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L., and Swanson, L. (1993). A method for severly constrained item selection in adaptive testing. Applied Psychological Measurement, 18, xxx-xxx.

Sympson J. B., and Hetter, R. D. (1985, October) Controlling item-exposure rates in computerized adaptive testing. Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Wainer, H. (1983). Pyramid Power: Searching for an error in test scoring with 830,000 helpers. The American Statistician, 37, 87-91.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., and Thissen, D. (1990). Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Table 1: Typical Flaws and Current Strategies

Flaw	Remove from scoring	Score all answers correct	Score more than one answer correct	Score with correct scoring guide
No correct answer	Yes	Yes	NA	NA
More than one correct answer	Yes	Yes	Yes	NA
Incorrect scoring guide	Yes	NA	NA	Yes

Table 2: Maximum observed exposure rates for various pools. Numbers in parentheses indicate exposure rates after additional extended Sympson and Hetter iterations.

	Discrete items	Reading Passages	Items for Passages
Baseline pool	.24	.19	.19
25 most popular items removed	.43 (.23)	.25 (.18)	.18 (.18)
2 most popular items removed	.48 (.23)	.22 (.20)	.19 (.20)
2 typical items removed	.24	.19	.21
2 most popular as first items removed	.26	.20	.19
2 most popular as last items removed	.29 (.24)	.20 (.20)	.19 (.20)

Table 3a: The 25 most popular items simulated as flawed.

	Unflawed simulation	Flawed simulation	Remove item[1]	All correct[1]	Correct key[1]	Appropriate[1]	Reduced simulation[1]
Mean score difference[2]	-	-	5.06	7.27	5.06	5.03	-
Proportion with flaws[3]	-	-	1.00	1.00	1.00	1.00	-
Mean test length[3]	30	30	25.2	30	30	27.4	30
Reliability[3]	.902	.757	.849	.844	.882	.856	.896

Table 3b: The two most popular items simulated as flawed.

	Unflawed simulation	Flawed simulation	Remove item[1]	All correct[1]	Correct key[1]	Appropriate[1]	Reduced simulation[1]
Mean score difference[2]	-	-	.44	1.25	.43	.40	-
Proportion with flaws[3]	-	-	.40	.40	.40	.40	-
Mean test length[3]	30	30	29.5	30	30	29.8	30
Reliability[3]	.902	.900	.900	.897	.899	.900	.904

[1] Methods of rescoring (see text).
[2] For simulees from a typical population who received flawed items.
[3] Weighted to reflect a typical population of test-takers.

Table 3c: Two typical items simulated as flawed.

	Unflawed simulation	Flawed simulation	Remove item[1]	All correct[1]	Correct key[1]	Appropriate[1]	Reduced simulation[1]
Mean score difference[2]	-	-	.51	1.15	.58	.57	-
Proportion with flaws[3]	-	-	.21	.21	.21	.21	-
Mean test length[3]	30	30	29.8	30	30	29.9	30
Reliability[3]	.902	.901	.901	.900	.901	.901	.899

Table 3d: The two items appearing most frequently in the first position simulated as flawed.

	Unflawed simulation	Flawed simulation	Remove item[1]	All correct[1]	Correct key[1]	Appropriate[1]	Reduced simulation[2]
Mean score difference[2]	-	-	1.76	2.19	1.76	1.74	-
Proportion with flaws[3]	-	-	.37	.37	.37	.37	-
Mean test length[3]	30	30	29.6	30	30	29.8	30
Reliability[3]	.902	.882	.896	.896	.898	.897	.897

[1] Methods of rescoring (see text).
[2] For simulees from a typical population who received flawed items.
[3] Weighted to reflect a typical population of test-takers.

Table 3e: The two items appearing most frequently in the last position simulated as flawed.

	Unflawed simulation	Flawed simulation	Remove item[1]	All correct[1]	Correct key[1]	Appropriate[1]	Reduced simulation[1]
Mean score difference[2]	-	-	.33	.66	.34	.33	-
Proportion with flaws[3]	-	-	.31	.31	.31	.31	-
Mean test length[3]	30	30	29.6	30	30	29.8	30
Reliability[3]	.902	.901	.902	.902	.902	.902	.898

[1] Methods of rescoring (see text).
[2] For simulees from a typical population who received flawed items.
[3] Weighted to reflect a typical population of test-takers.

Table 4: Root mean squared score differences for simulees who received flawed items.

	Between two CATS from baseline pool	Between a rescored CAT and a CAT from the reduced pool[1]	Between two CATS from the reduced pool[2]
25 most popular items are flawed	4.66406	5.37912	4.83548
2 most popular items are flawed	4.72784	4.82967	4.68439
2 typical items are flawed	4.77073	4.48419	4.89044
2 items most popular in first position are flawed	4.66886	4.90770	4.80466
2 items most popular in last position are flawed	4.62279	4.70788	5.01267

[1] Adaptive tests were rescored taking into account the nature of the flawed item.
[2] Reduced pools are smaller (by the number of flawed items) than the baseline pool.

Table 5: Correlations between scores for simulees who received flawed items.

	Between a rescored CAT and a CAT from the reduced pool[1]	Between two CATS from the reduced pool[2]
25 most popular items are flawed	.875	.896
2 most popular items are flawed	.866	.874
2 typical items are flawed	.848	.830
2 items most popular in first position are flawed	.894	.899
2 items most popular in last position are flawed	.777	.753

[1] Adaptive tests were rescored taking into account the nature of the flawed item.

[2] Reduced pools are smaller (by the number of flawed items) than the baseline pool.

Table 6: Root mean squared score differences for simulees who received flawed items from repeated simulations of two typical items as flawed.

	Between two CATS from baseline pool	Between a rescored CAT and a CAT from the reduced pool[1]	Between two CATS from the reduced pool[2]
Original	4.77073	4.48419	4.89044
Replication A	4.78349	4.82228	4.88737
Replication B	4.84196	4.92298	4.94335
Replication C	4.80101	4.87479	4.93549
Replication D	4.80059	4.81467	4.91804
Replication E	4.82087	4.45249	4.96439
Replication F	4.80337	4.77628	4.91934
Replication G	4.80061	5.10866	4.90923
Mean	4.80395	4.78207	4.92096
Standard Deviation	.02290	.21833	.02621

[1] Adaptive tests were rescored taking into account the nature of the flawed item.
[2] Reduced pools are smaller (by the number of flawed items) than the baseline pool.